PROCEEDINGS OF TDWG

Biodiversity Information Science and Standards 1: e20198
doi: 10.3897/tdwgproceedings.1.20198

OPEN ACCESS

Conference Abstract

Setting up an Interdisciplinary Data Infrastructure: Why Cooperation between Domain Experts and Computer Scientists Matters - An Experience Report from the GFBio Project

Birgitta König-Ries[‡], Dagmar Triebel[§], Robert Huber[|], Falko Glöckler[¶], Anton Güntsch[#], Janine Felden[|], Felicitas Löffler[‡], Jana Hoffmann[¶]

‡ Friedrich-Schiller-Universität Jena, Jena, Germany
§ Botanische Staatssammlung München, Munich, Germany
| MARUM, Universität Bremen, Bremen, Germany
¶ Museum für Naturkunde Berlin, Berlin, Germany
# Freie Universität Berlin, Berlin, Germany

Corresponding author: Birgitta König-Ries (birgitta.koenig-ries@uni-jena.de), Dagmar Triebel (triebel@bsm.mwn.de), Falko Glöckler (falko.gloeckler@mfn-berlin.de)

Received: 11 Aug 2017 | Published: 11 Aug 2017

Citation: König-Ries B, Triebel D, Huber R, Glöckler F, Güntsch A, Felden J, Löffler F, Hoffmann J (2017) Setting up an Interdisciplinary Data Infrastructure: Why Cooperation between Domain Experts and Computer Scientists Matters - An Experience Report from the GFBio Project. Proceedings of TDWG 1: e20198.
https://doi.org/10.3897/tdwgproceedings.1.20198

Abstract

The German Federation for Biological Data (GFBio; Diepenbroek et al. 2014) is implementing a national infrastructure for the preservation, integration, and publication of biological data collected in German research projects. GFBio is built upon an archive infrastructure comprised of nine data centers including PANGAEA and the major German Natural Science Collections (German Federation for Biological Data (GFBio) 2017a). Creating and running GFBio requires close collaborations within a highly interdisciplinary consortium. Bringing together expertise from collections, scientists in the relevant fields, biodiversity informaticians and computer scientists proved to be essential for designing and building this system.

© König-Ries B et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

GFBio is currently in its second funding phase. Essential services, required for the operation of the future infrastructure, have been successfully implemented. The realized technologies and tools use globally accepted standards as well as innovative concepts e.g., for data visualisation or semantic integration.

A portal (https://www.gfbio.org) provides a common point of access to all GFBio services: data submission, data discovery, data visualisation and analysis, a terminology service, and a help desk. In addition, archived research data is shared with international information infrastructures such as the Global Biodiversity Information Facility (GBIF) and the Biological Collection Access Service (BioCASE).

As the data centers use different systems and thus internally build upon different data structures (German Federation for Biological Data (GFBio) 2017b), the search functionality integrated in the portal is an good example of the collaboration between teams of different expertise.

Since the aim was to provide an integrated, faceted search, it was necessary to agree on common fields that can be used to feed the facets. Therefore, the GFBio data centers agreed on using ABCD 2.06 (Access to Biological Collection Data) as a common standard and specified thirty elements for data exchange. Here, it was essential to bring together (1) domain experts for defining which facets they consider useful for an effective search, (2) computer scientists for providing the implementation based on Elasticsearch (Elasticsearch 2017), (3) biodiversity informaticians for defining mappings between different standards and (4) data curators from the GFBio data centers and long-term repositories for negotiating the set of mandatory fields. The starting point for broader research data management workflows was derived from high-quality data provided via publishing pipelines established at each data center.

With that, primary collection and research data are available with metadata and data units according to the ABCD community standard and are ready to be reused following the FAIR data principles (Wilkinson et al. 2016): **F**indable, **A**ccessible, **I**nteroperable, **R**e-usable. Consequently, interdisciplinary cooperation is the GFBio data portal's measure of success.

Keywords

data curation, data management, research data, data archiving, FAIR principles, data portal, data standards, Access to Biological Collection Data, GFBio

Presenting author

Felicitas Löffler, Jana Hoffmann

References

- Diepenbroek M, Glöckner F, Grobe P, Güntsch A, Huber R, König-Ries B, Kostadinov I, Nieschulze J, Seeger B, Tolksdorf R, Triebel D (2014) Towards an Integrated Biodiversity and Ecological Research Data Management and Archiving Platform: The German Federation for the Curation of Biological Data (GFBio). In: Plödereder E, Grunske L, Schneider E, Ull D (Eds) Informatik 2014 – Big Data Komplexität meistern. GI-Edition: Lecture Notes in Informatics (LNI) – Proceedings., 232. Köllen Verlag, Bonn, 1711-1724 pp.
- Elasticsearch (2017) Elasticsearch - RESTful search and analytics engine. https://www.elastic.co/products/elasticsearch. Accessed on: 2017-8-08.
- German Federation for Biological Data (GFBio) (2017a) GFBio Data Centers. https://www.gfbio.org/about/data-centers. Accessed on: 2017-7-21.
- German Federation for Biological Data (GFBio) (2017b) Data exchange standards, protocols and formats relevant for the collection data domain within the GFBio network. https://gfbio.biowikifarm.net/wiki/Data_exchange_standards,_protocols_and_formats_relevant_for_the_collection_data_domain_within . Accessed on: 2017-7-21.
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, Silva Santos LBd, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, Schaik Rv, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, der Lei Jv, Mulligen Ev, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 160018. https://doi.org/10.1038/sdata.2016.18